DOCUMENT RESUME

ED 247 274 TM 840 448

AUTHOR Schratz, Mary K.
TITLE Vertical Equating: An Empirical Study of the Consistency of Thurstone and Rasch Model Approaches.
PUB DATE Apr 84
NOTE 15p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 24-26, 1984).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Achievement Tests; Elementary Secondary Education; *Equated Scores; Latent Trait Theory; Reliability; *Scaling
IDENTIFIERS *Rasch Model; Stanford Achievement Tests; Thurstone Scales; *Vertical Equating

ABSTRACT

To explore the appropriateness of the Rasch model for the vertical equating of a multi-level, multi-form achievement test series, both the Rasch model and the traditional Thurstone procedures were applied to the Listening Comprehension subtest scores of the Stanford Achievement Test. Two adjacent levels of these tests were administered in 1981 to a national student sample as part of the National Standardization Program. Pearson product-moment correlations and the conditional p-value procedure (Rentz) were used to further analyze the resulting four scaled scores. These analyses showed the consistency of both the Rasch and Thurstone models when applied to a vertical equating problem. Groups equal in ability were identified from equivalent scores derived through either equating method. There were, however, different patterns to the standardized differences between Thurstone scaled scores and between Rasch scaled scores obtained on the Primary One and Primary Three levels by fourth graders. There was a consistent pattern of departure found in the Thurstone data but not in the Rasch scaled scores. Both approaches have pitfalls in their application to vertical equating problems, but the Rasch model compared well to the traditional model. This has encouraging implications for computerized adaptive testing and customized test development and scoring. (BS)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED247274

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

M. K. Schratz

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Vertical Equating: An Empirical Study of the Consistency of Thurstone and Rasch Model Approaches

Mary K. Schratz

Navy Personnel Research and Development Center

Paper Presented at the Annual Meeting of the National Council on Measurement in Education
New Orleans, 1984

There are several reasons why achievement test series, administered in the traditional paper and pencil format, are often developed to consist of multiple levels and multiple parallel forms of each level. Test series of this kind are built to measure achievement across a wide grade range, making it necessary to construct many levels of the test that differ intentionally in the range of difficulty of test items and often in the specifications of its content. Parallel forms of a given level of the test are constructed to meet many practitioner needs. While parallel forms are designed to assess the same content and have comparable ranges of item difficulties, they often do differ somewhat in average difficulty. These considerations make it necessary to establish equivalent scores among different levels and forms of a test in order to compare performance on them.

Determining the relationship between raw scores on different levels of a test is a first step in building a developmental score scale that is continuous across all test levels. Multi-level achievement test series have traditionally been equated through the equipercentile method and then scaled through Thurstone's Absolute Scaling Technique (1925, 1927).

The scale that results from the application of these procedures is one which is dependent on the ability distribution of the sample on which it is based. In addition to assuming unidimensionality of the ability measured, this technique assumes that ability is normally distributed within age groups. Inadequate sampling of intended populations will produce inappropriate estimated scale values; samples which differ in their ability distributions will produce scale values different in magnitude and dispersion (Wright & Stone, 1979).

The limits of this method of scale construction have been recognized by a number of prominent measurement specialists (Angoff, 1960, 1962; Gardner, 1952; Gullicksen, 1950; Lindquist, 1953; Loevinger, 1947; Tucker, 1953). Gardner, for example, asserts that two desirable properties of a scale are: (1) that it is invariant with respect to the sample of items tested; and (2) that it is invariant with respect to the population used in its derivation.

That a different approach to scale construction may meet these measurement goals, yielding better results for the vertical equating problem, has been the hope of many experts. The conceptual appeal of the Rasch Model (or other latent trait theories) for this particular problem lies in the prediction that (1) items parameters that are invariant for different groups of persons can be estimated, and that (2) once a set of items has been calibrated, a given individual would receive the same measurement (estimated ability, except for errors of measurement), regardless of the particular subset of items taken. Slinde and Linn (1978) have pointed out that these properties "are precisely what is needed to solve the vertical equating problem" (p. 24).

The purpose of the present study was to explore the appropriateness of the Rasch Model for the vertical equating of a multi-level, multi-form achievement test series. The approach was to use

standard Rasch Model procedures for vertical equating as well as the traditional, Thurstone procedures for this problem.

METHOD

In the fall of 1981, the National Standardization Program for Stanford Early School Achievement Test, Stanford Achievement Test, and Stanford Test of Academic Skills was conducted. This research program entailed the testing of a nationally representative sample of more than 200,000 school children enrolled in kindergarten through the twelfth grade. One of the goals of this program was to obtain data for equating the ten adjacent levels of the Stanford series. In order to achieve this goal, a sample of more than 17,000 students was tested. Each student participating in this aspect of the standardization program was administered two adjacent levels of a single content domain of the series.

The Listening Comprehension subtest of Stanford Achievement Test was used in the analyses conducted in the present study. This subtest spans six levels of the test series. Of all the subtests included in the series, performance in listening is likely to be least dependent on the introduction of specific skills in the curriculum. The listening subtest was chosen for evaluating the equatings in an effort to avoid difficulties reported in past research (Loyd & Hoover, 1980; Lenke, 1981) when tests requiring the application of skills and knowledge that students have not had the opportunity to acquire have been equated. The content breakdown of this subtest is the same across all six levels. Items were selected for inclusion in the final edition of Stanford with consideration given to both traditional test construction criteria and fit to the Rasch Model. The number of students tested with each adjacent level pair of the Listening test is shown in Table 1.

Rasch Model procedures for scaling and equating tests and traditional, Thurstone scaling techniques were applied to the above described data base. The application of these procedures resulted in two raw score-to-scaled score conversion tables for each level of the Listening subtest, one for the Thurstone Model and one for the Rasch Model. Individuals' raw scores on the two adjacent levels of the test taken were converted to scaled scores. This resulted in four scaled scores for a given individual: a Thurstone scaled score and a Rasch scaled score at each of two adjacent levels.

In order to explore the similarity of the scorings of individuals on two different tests through the same scaling method, Pearson product-moment correlations were computed. For each grade tested, the correlation between scaled scores of the same type, obtained on adjacent levels, was determined.

Correlations provide valuable information on the orderings of individuals through each scale. Further analyses were conducted, however, to examine differences in measured achievement of individuals on adjacent test levels. Standardized differences were computed between scaled scores of each type for each individual and distributed by

grade. These distributions were visually examined and interpretations made with respect to the nature of departure from normality.

The accuracy of equating through each psychometric method was further evaluated by comparing mean item-performance data of examinees with equivalent scores on each test in an adjacent-level pair. The approach used has been referred to as the Conditional p-value Procedure by Rentz and has been described with respect to Stanford Achievement Test by Lenke and Rentz (1982). The basic approach is to consider equatings accurate if the two groups of individuals with equivalent scores attain equal p values on the same set of items. The p values for examinees with equivalent scores are referred to as conditional p values since they are conditioned on ability.

This procedure was applied in the following way. For each adjacent-level pair, those individuals scoring at the 10th, 25th, 50th, 75th, and 90th percentiles (in terms of raw score) on the lower test level were identified. Next, those individuals attaining the Rasch equated scores and equipercentile equated scores on the upper level corresponding to these five raw score points on the lower level were identified. Thus, for each adjacent-level pair, a total of fifteen score groups were identified. For each of these groups, p values were then computed for a set of randomly selected items drawn from both the upper and lower test levels. A mean p value for these items was computed for each of the fifteen groups. At all five levels of score performance, comparisons were then made between the mean p values earned by the Rasch equated-score groups on the upper and lower test levels and between the equipercentile equated-score groups on these same levels.

RESULTS AND DISCUSSION

Many of the analyses conducted in the present study provided evidence for the consistency of both the Rasch Model and the Thurstone Model in their application to a vertical equating problem. High positive correlations were found between scaled scores earned by the same individuals on adjacent test levels equated through the Rasch Model. The resulting correlations for each of the five adjacent-level pairs appear in Table 2. Correlations of the same order were found between Thurstone scaled scores and appear in Table 3. These results suggest that similarity of scoring on adjacent levels is achieved through either the Thurstone or Rasch Model approach to scaled score development.

Comparisons of mean item-performance data from the Rasch-equated score groups on adjacent levels and from the equipercentile equated-score groups through the Conditional p-value Procedure also yielded findings supportive of the adequacy of either model for equating tests of differing difficulty. For all adjacent level pairs, inspection of the mean p values obtained by students scoring at selected percentile points on the lower test level showed them to be similar in magnitude to those obtained by the Rasch equated-score groups. These same results were found for the equipercentile equated-score groups. Results of these analyses appear in Table 4.

These findings suggest that groups equal in ability were identified from equivalent scores derived through either equating method.

While the above described analyses were generally supportive of the adequacy of either model in its application to the vertical equating problem, analyses of standardized differences between scaled scores of each type obtained on adjacent levels that were conducted in this study revealed differences that could not be accounted for by measurement error. Distributions of standardized differences between Thurstone scaled scores that resulted were, on the whole, quite similar to the normal curve. An example of these data appears in Figure 1. There was, however, a tendency for these distributions, more pronounced at the lowest three grades, to be negatively skewed. It appears to have been somewhat easier for students in the early elementary school grades to achieve slightly higher scaled scores taking a test one level below that which is grade appropriate.

For the Rasch scaled scores, a different pattern of departure from normality was observed. For the same adjacent level pair shown in Figure 1 above, Rasch Model data appears in Figure 2. There was tendency for distributions of standardized differences to be more platykurtic that expected. Fewer concordant scaled scores, then, were obtained on adjacent levels than was anticipated. There was, however, no consistent trend across grades depicting the assignment of higher or lower scaled scores to students on the basis of easier or harder test level administrations, as was found for the Thurstone model. Such a tendency could potentially cause difficulties for test consumers assessing progress in school from one year to the next, if different test levels are administered. The results of the analyses of standardized differences found for each model do suggest, however, that neither approach in without pitfalls in its application to vertical equating problems.

There are several reasons why, in theory and in practice, traditional methods of scaling and equating are less desirable to apply to vertical equating situations than the Rasch Model. The results of the present study, in which the Rasch Model compared well to the traditional model in a vertical equating situation, are encouraging and carry implications for test development, many of which are currently being realized in the form of computerized adaptive testing and customized test development and scoring. Replication of the results of the present study, by applying the scaled score system developed in this study to a new, independent sample, and, perhaps, by conducting these analyses in other content areas would be helpful.

REFERENCES

Angoff, W. H. Measurement and scaling. In C. W. Harris (Ed.), Encyclopedia of educational research. (3rd ed.) New York: Macmillan, 1960. pp. 807-817.

Angoff, W. H. Scales with nonmeaningful origins and units of measurement. Symposium: Standard scores for aptitude and achievement tests. Educational and Psychological Measurement, 1962, 22, pp. 27-34.

Gardner, E. F. The importance of reference groups in scaling procedure. In A. Anastasi (Ed.), Testing problems in perspective. Washington, DC: American Council on Education, 1966. pp. 272-280.

Gulliksen, H. Theory of mental tests. New York: Wiley, 1950.

Lenke, J. M. A look at the outcomes of test equating using Rasch Model procedures. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Los Angeles, 1981.

Lenke, J. M. & Rentz, R. R. The use of the Rasch Model in the development of Stanford Achievement Test. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York, 1982.

Lindquist, E. F. Selecting appropriate score scales for tests. In Proceedings of the 1952 Invitational Conference on Testing Problems. Princeton, N. J.: Educational Testing Service, 1953, pp. 34-40.

Loevinger, J. A systematic approach to the construction and evaluation of tests of ability. Psychological Monographs, 1947, 61, pp. 1-47.

Loyd, B. H. & Hoover, H. D. Vertical equating using the Rasch Model. Journal of Educational Measurement, 1980, 17, pp. 179-193.

Slinde, J. A. & Linn, R. L. An exploration of the adequacy of the Rasch Model for the problem of vertical equating. Journal of Educational Measurement, 1978, 15, pp. 23-32.

Thurstone, L. L. A method of scaling psychological and educational tests. Journal of Educational Psychology, 1925, 16, pp. 433-451.

Thurstone, L. L. The unit of measurement in educational scales. Journal of Educational Psychology, 1927, 18, pp. 505-524.

Tucker, L. Scales minimizing the importance of reference groups. In Proceedings of the 1952 Invitational Conference on Testing Problems. Princeton, N. J.: Educational Testing Service, 1953, Pp. 22-28.

Wright, B. D. & Stone, M. H. Best test design: A handbook for Rasch Measurement Chicago: MESA, 1979.

TABLE 1

NUMBERS OF STUDENTS TESTED WITH TWO ADJACENT LEVELS OF THE LISTENING COMPREHENSION SUBTEST OF STANFORD ACHIEVEMENT TEST

Adjacent-Level Pair	Grade	Number of Students
Primary 1-Primary 2	3	968
Primary 2-Primary 3	4	857
Primary 3-Intermediate 1	5	808
Intermediate 1-Intermediate 2	6	993
Intermediate 2-Advanced	7	865
Intermediate 2-Advanced	8	704

TABLE 2

PEARSON PRODUCT-MOMENT CORRELATIONS
BETWEEN RASCH SCALED SCORES
OBTAINED ON ADJACENT LEVELS
BY GRADE

	Prim.2	Prim.3	Int.1	Int.2	Adv.
Prim.1 Grade 3	.72				
Prim.2 Grade 4		.68			
Prim.3 Grade 5			.75		
Int.1 Grade 6				.78	
Int.2 Grade 7					.77
Int.2 Grade 8					.68

TABLE 3

PEARSON PRODUCT-MOMENT CORRELATIONS
BETWEEN THURSTONE SCALED SCORES
OBTAINED ON ADJACENT LEVELS
BY GRADE

	Prim.2	Prim.3	Int.1	Int.2	Adv.
Prim.1 Grade 3	.68				
Prim.2 Grade 4		.66			
Prim.3 Grade 5			.75		
Int.1 Grade 6				.79	
Int.2 Grade 7					.78
Int.2 Grade 8					.70

TABLE 4

RESULTS OF THE APPLICATION OF THE CONDITIONAL p-VALUE PROCEDURE TO EACH ADJACENT-LEVEL PAIR

Primary 1-Primary 2 Equating

	LOWER LEVEL			UPPER LEVEL EQUIPERCENTILE			UPPER LEVEL RASCH EQUATED		
%ile	Scr[a]	N	MN P[b]	Scr[a]	N	MN P[b]	Scr[a]	N	MN P[b]
10	15	24	.612	14	18	.656	13	31	.594
25	19	46	.724	18	45	.737	17	39	.706
50	23	87	.868	22	60	.883	22	60	.883
75	25	118	.919	25	76	.925	25	76	.925
90	26	120	.949	27	79	.953	26	77	.939

Primary 2-Primary 3 Equating

	LOWER LEVEL			UPPER LEVEL EQUIPERCENTILE			UPPER LEVEL RASCH EQUATED		
%ile	Scr[a]	N	MN P[b]	Scr[a]	N	MN P[b]	Scr[a]	N	MN P[b]
10	18	19	.592	17	22	.655	15	8	.512
25	21	48	.700	22	35	.730	20	22	.668
50	24	7[illegible]	.836	27	31	.806	25	38	.796
75	27		.883	31	47	.893	32	40	.906
90	28	87	.902	34	32	.906	34	32	.906

Primary 3-Intermediate 1 Equating

	LOWER LEVEL			UPPER LEVEL EQUIPERCENTILE			UPPER LEVEL RASCH EQUATED		
%ile	Scr[a]	N	MN P[b]	Scr[a]	N	MN P[b]	Scr[a]	N	MN P[b]
10	18	11	.532	18	25	.524	16	11	.377
25	24	21	.636	23	28	.682	22	25	.632
50	30	46	.805	28	49	.779	29	41	.793
75	33	70	.857	32	57	.860	32	57	.860
90	36	48	.900	35	44	.910	35	44	.910

Intermediate 1-Intermediate 2 Equating

	LOWER LEVEL			UPPER LEVEL EQUIPERCENTILE			UPPER LEVEL RASCH EQUATED		
%ile	Scr[a]	N	MN P[b]	Scr[a]	N	MN P[b]	Scr[a]	N	MN P[b]
10	18	15	.513	16	8	.444	15	9	.389
25	24	27	.646	22	30	.693	21	27	.663
50	30	52	.786	27	56	.798	27	56	.798
75	34	61	.867	31	48	.868	32	40	.874
90	36	61	.908	34	41	.885	34	41	.885

TABLE 4 (cont'd)

RESULTS OF THE APPLICATION OF THE CONDITIONAL p-VALUE PROCEDURE TO EACH ADJACENT-LEVEL PAIR

Intermediate 2-Advanced Equating

	LOWER LEVEL			UPPER LEVEL EQUIPERCENTILE			UPPER LEVEL RASCH EQUATED		
%ile	Scr[a]	N	MN P[b]	Scr[a]	N	MN p[b]	Scr[a]	N	MN P[b]
10	20	28	.552	17	28	.612	15	34	.503
25	25	55	.654	21	86	.653	20	47	.618
50	30	103	.768	26	84	.757	26	84	.757
75	33	128	.841	30	88	.851	30	88	.851
90	36	88	.894	33	81	.917	34	50	.907

[a]Indicates earned raw score.
[b]Indicates mean p value.

Figure 1. Standardized differences between Thurstone scaled scores obtained on the Primary 2 and Primary 3 levels by fourth graders.

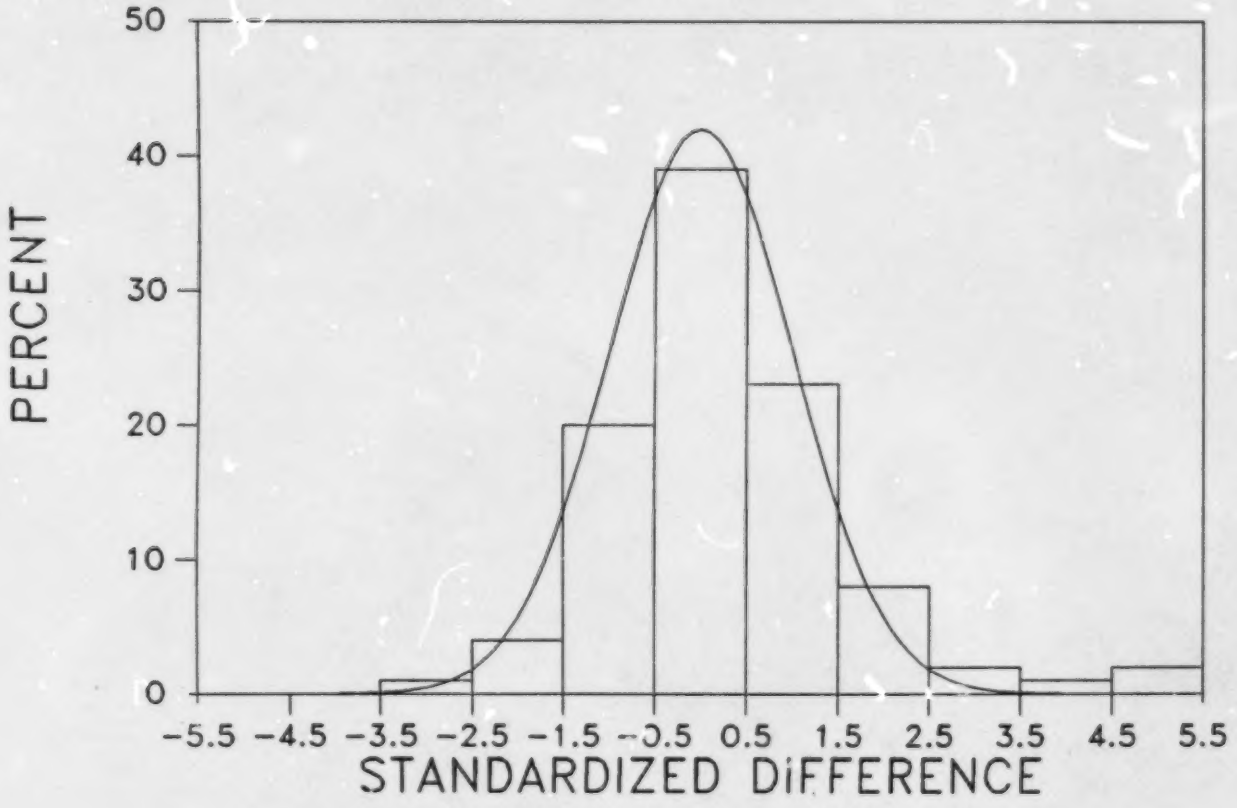

Figure 2. Standardized differences between Rasch scaled scores obtained on the Primary 2 and Primary 3 levels by fourth graders.